

MIMOR @ NTCIR 5: A Fusion-based Approach to Japanese Information Retrieval

Nina Kummer^{1,2}, Christa Womser-Hacker¹, Noriko Kando²

¹Universität Hildesheim, Germany

²National Institute of Informatics, Tokyo, Japan

nina@nii.ac.jp, womser@uni-hildesheim.de, kando@nii.ac.jp

Abstract

In our first participation in the NTCIR Workshop, we focused on the evaluation of the relative effectiveness of different indexing approaches (word-based, N-gram-based, and yomi- or pronunciation-based) for Japanese IR and the benefits of their fusion. Our MIMOR (“multiple indexing for method-object relations in IR”) system has already proved very effective in CLEF¹. The results show that our approach is also promising for Japanese IR.

Keywords: Japanese IR, fusion, yomi-based indexing.

1 Introduction

The department of Information Science of the University of Hildesheim, Germany, has been participating regularly in the CLEF workshops with its own information retrieval system MIMOR since 2002 [1]. The system supports most European languages. MIMOR is modeled as an open information retrieval system designed to combine different approaches in information retrieval within one Meta system. This allows for the exploration of the performance of individual retrieval devices and/or approaches on the one hand, and also profits from the advantages of the best-performing technologies for an optimal retrieval result. These characteristics should also be utilized when tackling Japanese IR.

2 Background

2.1 Challenges in Japanese IR

The main challenge for IR systems working with Japanese documents lies in the tokenization step.

With Japanese lacking explicit boundaries between words in a sentence, indexing procedures are quite different from those used for European languages. Therefore, we have put our focus on different indexing strategies for Japanese.

Another challenge in Japanese IR is orthographic variety. It is a very frequent phenomenon owing to the combined usage of four different scripts within one writing system (kanji, hiragana, katakana, and Roman characters). The most common forms of orthographic varieties comprise cross-script variants (words which can be represented in different scripts), okurigana variants (differing in the number of syllables expressed in hiragana in addition to a kanji stem), hiragana variants (irregularities in the use of hiragana), kanji variants, phonetic substitutes, and katakana variants. A comprehensive overview of the types of orthographic variety can be found in Halpern [2, 7]. From the information retrieval point of view, we can classify orthographic varieties in Japanese into two groups:

1. Variants originating from a different written representation of the same phoneme (cross-script variants, okurigana variants, hiragana variants, kanji variants, and phonetic substitutes).
2. Variants originating from a different interpretation of the sound structure to be represented (katakana variants).

Variants in the first group share the same pronunciation. This fact can be exploited for information retrieval, if the terms are matched using their pronunciation instead of their written representation. Variants of the second type only differ in minor aspects, i.e. usually only one character. We suppose that matching terms based on their editing distance may be an effective means of retrieving documents that contain katakana variants of a search term. The latter approach was tested in an earlier study [6], but not further pursued in the experiments described here.

2.2 Yomi-based Indexing

Yomi-, or pronunciation-based, indexing is not a new strategy for use in Japanese IR. In contrast, it

¹ <http://clef-campaign.org>

is a rather old technique, which used to be employed before the introduction of double-byte processing on computers. In those days, information processing systems used the katakana syllabary to represent Japanese text phonetically. The yomi-based index has been abandoned since the introduction of double-byte character handling, as the Japanese language is very rich in homophones and the kanji characters convey important information for disambiguation.

Although a yomi-based index may incur losses in precision through ambiguous homophones, we suppose that it may be valuable for the handling of orthographic varieties. The advantage of a pronunciation-based index is that it is insensitive to orthographic variants (e.g., okurigana, kanji, or kana variants), as it is independent from the written form of a word. Fusion with other index types can help to reduce the negative influence of ambiguous homophones.

2.3 Fusion in Japanese IR

Similarly to the findings in TREC, the evaluations of the NTCIR Workshop series have not produced one clearly superior system, but rather comparably well performing systems using completely different approaches.

The evaluations of the NTCIR Workshop series have not produced a clearly superior indexing approach, but rather, show systems performing equally well using very different indexing approaches. The two basic approaches are word-based indexing, which requires Natural Language Processing (NLP) techniques, and character N-gram indexing, which is language independent. Both strategies lead to similar results, but their effectiveness varies case-by-case [8, 12]. To take maximum advantage of the strengths of the individual approaches, while at the same time minimizing their disadvantages, a number of enhanced approaches have been suggested. Among these are the “combination-of-evidence”, or fusion approaches. These approaches merge the result lists obtained using more than one index type, usually by coupling word-based and N-gram-based indices. The results show that ranking documents based on a multiple index search is a promising strategy in Japanese information retrieval [3, 9, 11].

3 System Overview

3.1 The MIMOR Approach

Here, the meta system MIMOR is based on Lucene as basic retrieval engine. Lucene allows the combination of various indexing systems. For the overall system, MIMOR should profit from users’

relevance assessments in order to learn which combinations of object representations and information retrieval functionality lead to good performance. An internal evaluation procedure, which is realized via a blackboard model, permanently registers which resource produces good results and which one does not. Well-performing techniques gain high weights, poorly-performing ones are excluded over time.

3.2 MIMOR for Japanese

We created three different indices – an N-gram-based, a word-based, and a yomi-based index.

For the N-gram index, hiragana characters were discarded, katakana and roman character strings were left in their original form, and kanji character strings were divided in overlapping bigrams.

The morphological analysis for the word- and yomi-based indices was carried out with ChaSen². Out-of-vocabulary words, i.e. words not recognized by ChaSen, were divided into bigrams. This can be called a hybrid approach [4].

For the yomi-based index, in the case of more than one suggested readings for one term, the readings were indexed as separate terms (e.g. ナマモノ and セイブツ for 生物). The pronunciation was extracted from the ChaSen output.

The fusion strategy we adopted was Z-Score, which was successfully employed by Savoy [10] in the NTCIR-4 data, and yielded the best results in our earlier study [5].

Z-score fusion allows for a normalized linear combination of the search results. The contribution of the individual systems is controlled using a weight represented by the parameter α (see. Equation 1).

$$Z\text{-Score}RSV_i = \alpha \cdot \left[\frac{RSV_i - Mean^i}{Stdev^i} + \delta^i \right] \quad (1)$$

$$\delta^i = \frac{Mean^i - Stdev^i}{Stdev^i}$$

Equation 1. Z-Score.

Note: RSV stands for “Retrieval Status Value”, i.e., the score assigned to a retrieved document

3.3 Optimization Strategies

A stoplist for each individual index was created determining the 100 most frequent index terms and we decided heuristically which of those terms should be discarded. In the case of the scientific abstracts collection, we decided to discard terms such as 研究 (research), 方法 (method), 実験 (experiment), 検討 (investigation, study), 結果

² <http://chasen.aist-nara.ac.jp/hiki/ChaSen/>

(result), and 目的 (purpose), which act as structure words, and are to be found in practically every scientific document. Similarly, we discarded terms such as 記事 (article) and 問題 (problem) for queries within the news domain. The yomi stoplist contained some equivalents of typical stop terms that were also to be found in the word-based stop list, such as モノ (thing), as well as the numerals 0 (レイ、ゼロ) to 9 (キユウ), and a number of individual syllables.

Pseudo Relevance Feedback was carried out using the Robertson Selection Value for the selection of expansion terms. For each individual index, the optimal parameters (number of relevant documents to be retrieved and number of terms to be extracted) were determined beforehand using the NTCIR-4 test collection.

4 Experiments and Results

4.1 Submitted Runs

In former experiments using a part NTCIR-4 test collection (Mainichi Shinbun '98), we found the following order of performance of single indices: bigram index, yomi-based index, word-based index [6]. Although the differences were only minor (about 1% precision) and the characteristics of the document collections might differ, we used this result as a basis for our NTCIR-5 experiments and tested, beside the single yomi-based index, a combination of all three indices vs. a combination of only the yomi- and bigram-based indices.

The weights of the individual indices within the fusion experiments were set to 1. Table 1 shows the results of all submitted runs. Also included in our official runs are a Title-only and a Description-only run using a combination of the yomi- and bigram-based indices.

Combination	Fields	Relaxed	Rigid
Yomi	TDNC	.3823	.3105
Yomi+Bigram	TDNC	<u>.3952</u>	<u>.3169</u>
Yomi+Bigram+Word	TDNC	.3888	.3063
Yomi+Bigram	T	.2836	.2015
Yomi+Bigram	D	.2541	.2251

Table 1. MAP of individual submitted runs.

The combination of the yomi- and bigram-based index performed best, followed by the triple index. Interestingly, the yomi-based index alone already shows quite a high performance. The combination of the yomi- and word-based index reaches a significant improvement of 3.37% over the single yomi-based index (relaxed judgements, T-test, $p=0.05$). The Title- and Description-only runs

clearly performed worse. Figure 1 shows a comparison of the precision reached by the single yomi-based index compared to our fusion experiments.

4.2 Post-submission Runs

In order to determine the relative performance of the individual indices within our fusion approach, we carried out two more runs: single word-based index and single bigram-based index using all topic fields.

Index Type	Fields	Relaxed	Rigid
Yomi	TDNC	<u>.3823</u>	<u>.3105</u>
Word	TDNC	.3478	.2634
Bigram	TDNC	.3265	.2485

Table 2. MAP of single-index runs.

It turns out that the order of performance changed compared to our previous experiments with the Mainichi'98 collection. The yomi-based index performs significantly better than the word- and bigram-based indices (relaxed judgements, T-test, $p=0.05$).

Figure 2 shows a comparison of the precision reached by each individual index per topic. Figures 3 and 4 show the Recall-Precision curve and the Frozen Ranks graph for the three individual indices, respectively. In both cases, the yomi-based index clearly outperforms both word- and bigram-based index.

4.3 Topic Analysis

In a closer analysis of the topics with a clear performance gap between the different indices, we found two interesting cases:

Topic 18, which deals with “Tobacco business, accusation, compensation” shows a very low performance of the bigram-based index. It turns out that this is due to the central word たばこ (ta-ba-ko, tobacco in English), which is written in hiragana and therefore discarded in the bigram approach. Whereas most hiragana words are not content-bearing, たばこ is.

In topic 49, dealing with “wild animal, crops, damage”, the yomi-based indexing strategy showed much better performance than both word- and bigram-based approach. This can be traced back to the pronunciation of 作物, which may be サクブツ (sa-ku-bu-tsu), in the sense of “literary work“, or サクモツ (sa-ku-mo-tsu), meaning “crops”, and of the related compound 農作物, which stands for “crops” and can be pronounced ノウサクブツ (nou-sa-ku-bu-tsu) or ノウサクモツ (nou-sa-ku-mo-tsu). Having more than one pronunciation

boosts the weight of a term, as it will be indexed and searched for with all pronunciation alternatives. In cases where the semantics of a term change with its pronunciation, (as with sa-ku-bu-tsu – literary work vs. sa-ku-mo-tsu – crops), this may lead to confusion and reduced precision. In this case, however, the system rather profited from a boost of the important terms 作物 and 農作物 by a factor of 2.

Although these findings explain the performance differences between the individual indices, no general conclusions can be drawn from them, as they deal with particular cases. The term boosting due to several pronunciations is a rather random phenomenon and does not necessarily have a positive effect.

5 Conclusion

We successfully implemented Japanese language support into our IR system. The proposed yomi-based index, which works on the pronunciation of Japanese terms, showed an excellent performance. We were further able to improve retrieval effectiveness by a multiple index fusion approach. The optimal weights per individual index still need to be determined.

As the ultimate goal of MIMOR is the automatic selection of the optimal combination of retrieval methods per individual retrieval case, it would further be desirable to see if topic or document properties can be identified which lead to a higher performance of a certain index type. Ultimately, this would lead to the automatic adaptation of retrieval methods to each individual retrieval case.

6 References

- [1] Hackl, R.; Mandl, Th. & Womser-Hacker, Ch. (2005): Ad-hoc Mono- and Multilingual Retrieval Experiments at the University of Hildesheim. In: Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Sept. 2005, Vienna.
- [2] Halpern, J. (2002): Lexicon-Based Orthographic Disambiguation in CJK Intelligent Information Retrieval. In: Proceedings of the 19th Conference on Computational Linguistics, COLING-2002, August 24-September 1, 2002, Taipei, Taiwan.
- [3] Jones, G. J. F.; Sakai, T.; Kajiura, M. & Sumita, K. (1998): Experiments in Japanese Text Retrieval and Routing Using the NEAT System. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, pp. 197-205.
- [4] Chow, K. C. W., Luk, R. W. P., Wong, K. F. & Kwok, K. L. (2000): Hybrid term indexing for different IR models. In: Proceedings of the fifth international workshop on Information retrieval with Asian languages. Hong Kong, China, pp. 49-54.
- [5] Kummer, N.; Womser-Hacker, Ch. & Kando, N. (2005): Re-Examination of Japanese Indexing: Fusion of Word-, N-gram- and Yomi-Based Indices. In: Proceedings of the 11th Annual Meeting of The Association for Natural Language Processing, March 14-18, 2005, University of Kagawa, Kagawa Prefecture, Japan, pp. 221-224.
- [6] Kummer, N.; Womser-Hacker, Ch. & Kando, N. (2005): Handling Orthographic Varieties in Japanese IR: Fusion of Word-, N-gram-, and Yomi-Based Indices across Different Document Collections. AIRS 2005, October 13-15, 2005, Jeju Island, Korea, pp. 666-672.
- [7] Kummer, N.; Womser-Hacker, Ch. & Kando, N. (2005): Handling Orthographic Varieties in Japanese IR: Fusion of Word-, N-gram-, and Yomi-Based Indices across Different Document Collections. NII Technical Report, July 2005.
- [8] Ozawa, T.; Yamamoto, M.; Umemura, K. & Church, K. W. (1999): Japanese Word Segmentation Using Similarity Measure for IR. In: Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, August 30-September 1, 1999, Tokyo, Japan, pp. 89-96.
- [9] Sakai, T.; Shibazaki, Y.; Suzuki, M.; Kajiura, M.; Manabe, T. & Sumita, K. (1999): Cross-Language Information Retrieval for NTCIR at Toshiba. In: Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, August 30-September 1, 1999, Tokyo, Japan, pp. 137-144.
- [10] Savoy, J. (2004): Report on CLIR Task for the NTCIR-4 Evaluation Campaign. In: Proceedings of the Fourth NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, pp.178-185.
- [11] Vines, Ph. & Wilkinson, R. (1999): Experiments with Japanese Text Retrieval Using mg. In: Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, August 30-September 1, 1999, Tokyo, Japan, pp. 97-100.
- [12] Yoshioka, M.; Kuriyama, K. & Kando, N. (2002): Analysis of the Usage of Japanese Segmented Texts in NTCIR Workshop 2. In: Proceedings of the Second NTCIR Workshop on Research in Chinese and Japanese Text Retrieval and Text Summarization, National Institute of Informatics, Tokyo, Japan, pp. 291-296.

